Relations of Carrier and Side-Bands in Radio Transmission[1]

By R. V. L. HARTLEY

SYNOPSIS: This paper discusses generally the characteristics of carrier transmission as applied in radio and in carrier current communication over wires and analyses the factors which affect the faithfulness with which such systems reproduce the signals imparted to them. Modulation is shown to generate two side bands which, with respect to frequency, lie just above and just below the carrier frequency, the frequency width of each side band being the same as the frequency width of the original signals. Upon detection, currents of frequencies corresponding to the difference frequencies between all the possible pairs of component frequencies of the side bands and carrier are produced and, in general, are all found in the received message. It is therefore impossible to transmit messages, either telephone or telegraph, by carrier which will be absolutely free of distortion, but since the amplitude of any particular difference frequency is proportional to the product of the amplitudes of its two generating frequencies the distortion can be reduced below a troublesome value by maintaining the amplitude of the original carrier sufficiently large with respect to the amplitudes of the signal components. The distortion which arises from phase shifts between the component frequencies of the transmitted message and carrier is also considered.

The paper discusses single side-band transmission and carrier suppression with homodyne detection and their various merits are pointed out. Single side band transmission reduces the width of frequency band required for each message. Carrier suppression results in a saving of power, or a more economical expenditure of power, it having been determined that for proper freedom from distortion the power of the carrier component alone, when transmitted, should be rather larger than the peak power in a carrier suppression system. The use of local carrier in homodyne radio telephony assists in frequency selection in the same way as does the heterodyne wave in radio telegraph reception. The same applies also to static interference and, as the object of high power stations is to make the signals large compared with static, there is a gain in concentrating the power in side bands rather than in carrier.

Consideration of distortion arising from phase shifts shows that in homodyne telegraphy distortion can most readily be avoided by transmitting both side-bands, while in telephony these factors favor the transmission of only one side-band. The power of the reproduced signals is twice as great with two side-bands as with one, but there is no choice between one and two side-bands on the basis of the ratio of signals to interference.

The result of using a local detecting frequency which is not exactly equal to the original carrier frequency is discussed, and a balanced detector is described by means of which the distorting effect of the received carrier may be very much reduced. Considering a local carrier which is out of synchronism with the original carrier, it is again found that single side-band transmission is most favorable in telephony, and the transmission of both side-bands is best in telegraphy.—*Editor.*

AS indicated by the title, this paper will discuss some of the phenomena associated with radio transmission in terms of the carrier currents and side-bands into which a modulated wave may

[1] Presented before The Institute of Radio Engineers, New York, December 13, 1922. Printed in the Proceedings for February, 1923, and reprinted here by permission of the Institute.

be resolved. The use of these terms implies a point of view which perhaps is employed less commonly in radio engineering than in some of the other branches of the communication art. For this reason, I shall, at the risk of repeating much that is already in the literature,[2] review such of the fundamentals of this viewpoint as are necessary to an understanding of what is to follow.

Analysis of a Signal Wave

Briefly stated, the point of view is that any signaling wave may be resolved into sustained sinusoidal components, which may be thought of as traversing the system as individual currents and recombining at the receiving end to form the reproduced signal. The possibility of such a resolution has been demonstrated mathematically and the formulas for evaluating the amplitudes and phases of the components are well known. A periodic wave may be expressed as a Fourier series, that is, as the sum of an infinite series of components the frequencies of which may be thought of as harmonics of a fundamental frequency which is equal to the frequency of repetition of the wave. Such a resolution, however, is not directly applicable to the waves employed in communication, for by their very nature they are not periodic. A communication system must be capable of transmitting any individual symbol regardless of what precedes or follows it. We may, however, resolve such an aperiodic wave by the mathematical device of assuming it to be one cycle of a periodic wave in which the interval between successive occurrences of the disturbance in question approaches infinity. The frequency of repetition is then infinitesimal. The fundamental frequency of the Fourier series and the frequency interval between adjacent components are also infinitesimal; that is, the series of discrete lines of the Fourier series spectrum merge into a continuous spectrum. Mathematically this continuous spectrum is represented by the expression

$$F(t) = \int_o^\infty S \cos (qt + \theta)\, dq, \qquad (1)$$

which is known as the Fourier integral. Physically we are to picture this infiinte series of sustained sinusoids as having such amplitudes and phases that the algebraic sum of their instantaneous values is

[2] "Carrier Current Telephony and Telegraphy," by E. H. Colpitts and O. B. Blackwell; "Transactions of the American Institute of Electrical Engineers," volume XL, page 205, 1921. "Application to Radio of Wire Transmission Engineering," by L. Espenschied; presented before The Institute of Radio Engineers, January 23, 1922.

zero for all instants before and after the disturbance in question, and equal to the instantaneous value of the wave thruout its duration. In Fig. 1, curve A represents a telegraph dot, curve B gives the rela-

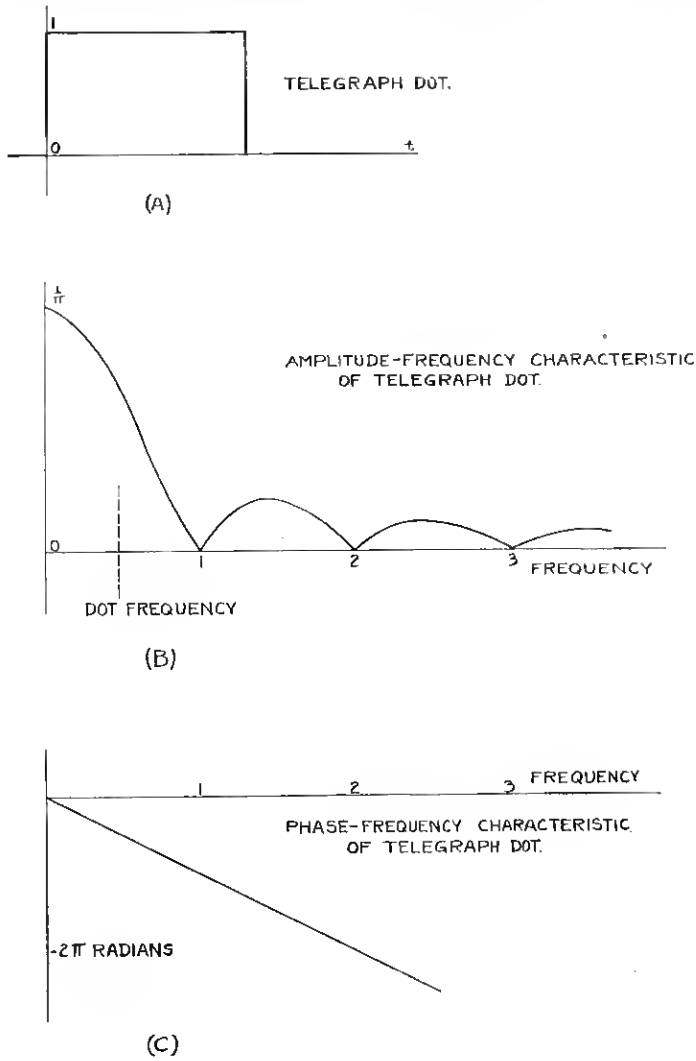

Fig. 1

tive amplitudes, S, of its components plotted against their frequencies, and curve C, their phase, θ, also as a function of frequency. The so-called "dot frequency" corresponding to a sustained succession of such dots is indicated on curve B.

It is obvious that if either the amplitudes or the phases of the components be distorted, their instantaneous sum will be changed; that is, the wave resulting from their re-combination will be a distorted reproduction of the original wave. Also, those parts of the frequency range in which the amplitude is negligibly small can contribute little to the reproduced wave, and the elimination of all components in those ranges will have little effect on the quality of reproduction. Just what ranges it is essential to retain depends upon the nature of the signal and the standard of reproduction that is set up. What is important for present purposes is the fact that the faithfulness with which a system will reproduce any arbitrary signal disturbance is deducible, in theory at least, from a knowledge of its transmission of sustained single frequencies. By this is meant a knowledge of how the relation, both in amplitude and phase, between the input and output sinusoidal wave varies as the frequency of the wave is progressively varied thruout the frequency range.

Analysis of a Modulated Wave

Let us assume now that a radio system is called on to transmit such a signal wave, $F(t)$, which may be either a telephone or a telegraph signal. If, as is commonly assumed, the modulator causes the amplitude of the carrier wave, $C \cos p\,t$, to be varied in accordance with the signal, the resulting modulated wave may be expressed as

$$m = C\,[1 + k\,F(t)]\cos pt, \tag{2}$$

where k is a factor which measures the so-called degree of modulation. If the largest negative value of $k\,F(t)$ is just equal to unity, so that the instantaneous amplitude of the carrier wave just falls to zero, the modulation is said to be complete. The significance of complete modulation will be discussed later.

Now let us resolve the signal wave into its infinite series of components, each of the form $S \cos (qt+\Theta)$, where S and Θ vary with the frequency $\frac{q}{2\pi}$. Neglecting non-essential frequencies, q may be considered to cover a range from q_1 to q_2. If this value of $F(t)$ be substituted in (2) we get

$$m = C \cos pt + k\,C \cos pt \int_{q1}^{q2} S \cos (qt+\Theta)\,dq. \tag{3}$$

The first term, which is independent of the signal, represents a component having the carrier frequency, $\frac{p}{2\pi}$. The second term represents an infinite series of terms each derived from only one component of

the signal. Hence each component of the signal is represented in the modulated wave by an expression of the form,

$$k\,C\,S\cos(qt+\theta)\ \cos pt=\tfrac{1}{2}k\,C\,S\ \{\cos[(p+q)t+\theta]+\cos[(p-q)t-\theta]\}. \quad (4)$$

This represents two sinusoidal components, the frequencies of which differ from that of the carrier by the frequency of the particular signal component. The similar expressions for the other signal components each yield a pair of components similarly placed with reference to the carrier. All of these taken together form a pair of spectra or frequency bands extending on either side from the carrier frequency in the same way that the spectrum of the signal extends from zero frequency. These bands of frequencies are spoken of as "side-bands" and the component currents of these frequencies as "side-band currents," or, more often, simply as "side-bands." The side-band which extends upward in frequency from the carrier is called the "upper side-band," and the other, which extends downward, the "lower side-band."

The form of these side-bands is shown schematically in Fig. 2, where purely arbitrary curves are used to represent the amplitudes and

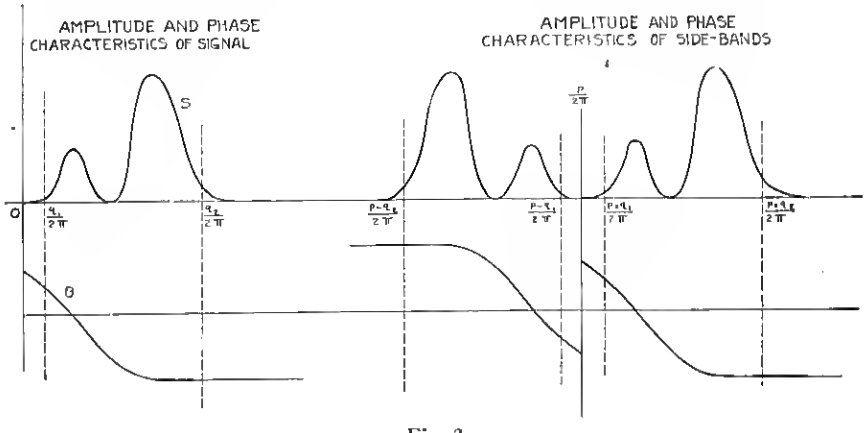

Fig. 2

phases of the signal components over a limited frequency range. It will be seen that the corresponding curves for the upper side-band are derived from these by displacing them along the frequency axis by the amount of the carrier frequency. The amplitude curve of the lower side-band is derived by inverting that of the upper with respect to the carrier frequency. For the phase curve of the lower side-band that of the upper is to be similarly inverted and also reversed in sign.

The actual magnitude of the side-band currents relative to the carrier depends on the degree of modulation, k, of equation (2). For commercial telephony the limits of the essential band may be taken roughly as 200 and 2,000 cycles. If high quality speech or music is to be transmitted, a wider band is required. For telegraphy the band width required varies widely with the speed of sending and the type of apparatus used. In general, it is desirable to preserve very low frequencies, which means that the two side-bands practically meet at the carrier frequency.

Reproduction of the Signal Wave

Having arrived at a picture of the modulated wave as given by equation (4), we shall first discuss the reproduction of the signal from this as it stands, and then consider the effect on this reproduction of various modifications to which the modulated wave may be subjected before or during the process of detection. While any device in which the current-voltage characteristic is non-linear may be used as a detector, the operation of the vacuum tube lends itself to analysis because of its approximation to a parabolic current-voltage relation. That is, we may write,

$$i = a_0 + a_1 v + a_2 v^2, \tag{5}$$

where v is the voltage impressed on the grid, in this case the modulated wave, and i is the resulting current. As the first term is independent of v and the second represents simple amplification, detection [3] can result only from the third term, $a_2 v^2$. Since a_2 multiplies all components of v^2 alike, we may neglect it and simply consider the square of the expression for the modulated wave. This results in a series of terms which are the squares of the individual components and another which are their products taken in pairs. Since

$$\cos^2 x = \tfrac{1}{2}(1 + \cos 2x), \tag{6}$$

the square terms will yield only direct current, and currents of approximately twice the carrier frequency. The product terms, each of which contains the product of two cosines, may, as in the case of the modulated wave above, be transformed into the sum of two cosine terms the frequencies of which are respectively the sum and difference of the component frequencies. Of these only the difference frequencies can lie in the range of the original signal. In other words, we may think of the reproduced wave as made up of the sum of all the

[3] In practice this parabolic law seldom holds strictly, and secondary contributions are made to the detected wave by terms of higher power.

heterodyne beat notes resulting from all the pairs of component sinusoids of the modulated wave.

The carrier component, $C \cos pt$, beating with a component of the upper side-hand, $\frac{1}{2} k C S \cos [(p+q) t + \theta]$, equation (4), gives the beat note or reproduced component,

$$r_+ = \tfrac{1}{2} k C^2 S \cos (qt+\theta), \tag{7}$$

which is identical in frequency and phase with the corresponding component of the signal, and has an amplitude proportional to that of the signal component. Exactly the same expression results from beating the carrier and the corresponding component of the lower side-band. These two low frequency components, being in phase, add directly to give

$$r = k C^2 S \cos (qt+\theta) \tag{8}$$

as the reproduced component. As the factor $k C^2$ is independent of q, all of the signal components are reproduced with the same relative amplitudes and phases, as in the original signal. Their sum is therefore $k C^2 F(t)$, and the signal is accurately reproduced.

However, there are still other components of the modulated wave to be considered. Every pair of components in one side-band beat to give the difference of their frequencies, which is also the difference of the corresponding signal components. The corresponding pair of components of the other side-band yield an identical component and the two add in phase. Similarly every component of one side-band beats with every component of the other, giving in each case the sum of two component frequencies of the signal wave. Like the difference frequencies, each of these sum frequencies is produced twice. The combination of the components of the two side-bands which were derived from the same signal component yields a component of twice the frequency of the signal component. The addition of these extraneous components serves to distort the reproduced wave in a manner quite similar to that of external interference. It is of interest therefore to consider the magnitude of these distorting components relative to the reproduced signal. The product of two side-band components of amplitudes $\frac{1}{2} k C S$ and $\frac{1}{2} k C S'$, equation (4), gives as the amplitude of one of the two components of the difference frequency, $\frac{q-q'}{2\pi}$, $\frac{1}{4} k^2 C^2 S S'$. Comparing this with the amplitude, $\frac{1}{2} k C^2 S$, equation (7), of one of the two reproduced signal components of frequency $\frac{q}{2\pi}$, the ratio of the undesired to desired component is found to be $\frac{1}{2} k S'$. It is evident that this type of distortion

increases with the degree of modulation, k, or, as will be discussed more fully later, with the ratio of carrier to side-band.[4]

Single Side-Band Transmission

So far it has been assumed that the wave applied to the detector is identical with that produced by the modulator, a condition seldom encountered in practice. For, in addition to the undesired modifications which the modulated wave undergoes because the transmission characteristics of practical circuits are not ideal, there are other changes which when properly made yield distinct advantages. These intentional changes will be discussed first.

It will be remembered that any component of the signal can be reproduced by the combination of the carrier with either side-band. Hence it is unnecessary to transmit both side-bands. Suitably designed electrical filters make it possible to transmit one side-band and effectively suppress the other.[5] This makes possible a very great saving in the frequency range required per channel. It is of particular importance for long wave radio telephone transmission where the width of a single side-band is so large a fraction of the total frequency range available that the number of independent channels is at best very limited. The intensive development of a limited frequency range by the use of single side-band transmission has probably progressed farthest in connection with carrier telephony over wires. Here commercial service is being given over circuits on which the carrier currents of adjacent channels are separated by only 3,000 cycles. It is obvious that the transmission of both bands would nearly double this separation, thereby halving the number of channels per circuit. There is, of course, no reason why similar savings may not be effected in the field of radio transmission. In addition to this major advantage there is an incidental improvement in the quality of reproduction, for the distorted components resulting from beats between components of the two side-bands, that is, the sums of the signal frequencies, are eliminated.

Carrier Suppression and Homodyne Reception

The other important modification has to do with the so-called " unmodulated " component of carrier frequency, $C \cos p\,t$, in equation

[4] A similar form of distortion generally occurs in modulation, resulting in new components being produced in the frequency range of the side-band.

[5] For a description of such filters see the Colpitts and Blackwell paper referred to above.

(3). As already pointed out, good signal reproduction requires that at the detector this shall not be too small relative to the side-bands. However, it is merely a continuous alternating current, and does not itself partake of the signal variations. It is therefore immaterial whether it is transmitted from the modulator or is supplied to the detector by a local source such as an oscillator. The elimination of this component from the modulated wave at the sending station is spoken of as "carrier suppression," and its re-introduction at the receiving end as "homodyne" or "zero beat" reception. The term homodyne implies supplying the same wave as distinguished from heterodyne, meaning another. Zero beat refers to the bringing of the local carrier into synchronism with the sending carrier by reducing the beat note between them to zero frequency. While homodyne reception is essential to carrier suppression, the reverse is not true. The reception of an ordinary modulated wave may sometimes be improved by the addition of carrier at the receiving end.

The primary advantage of carrier suppression lies in the saving of sending power which it makes possible, or, what is equivalent, the increase in range made possible when all the power of a given station is utilized in the side-band. Of the various ways in which this suppression may be accomplished, the simplest is by the use of a so-called balanced modulator as shown schematically in Fig. 3. Carrier fre-

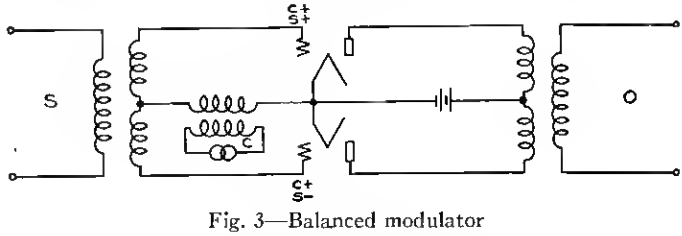

Fig. 3—Balanced modulator

quency from the source *C* is applied to the grids of two vacuum tubes in the same phase, while signal currents, indicated at *S*, are applied to the two in opposite phase. The two plate circuits are differentially connected with a common output circuit. In the absence of signaling current the amplified carrier frequency currents from the two tubes neutralize each other and nothing is transmitted. With the application of signaling current one grid is raised in potential and the other lowered, with the result that more radio frequency is developed by the first tube than by the second and the excess appears at O. The magnitude of this radio frequency current is proportional to the instantaneous value of the signaling current. Upon reversal of the

direction of the signaling current the effect of the second tube predominates and radio frequency is again transmitted, this time with the phase reversed, owing to the differential connection. The wave form of a signaling current and the resulting output current are roughly as shown in Fig. 4, *A* and *B*. If this output be amplified for trans-

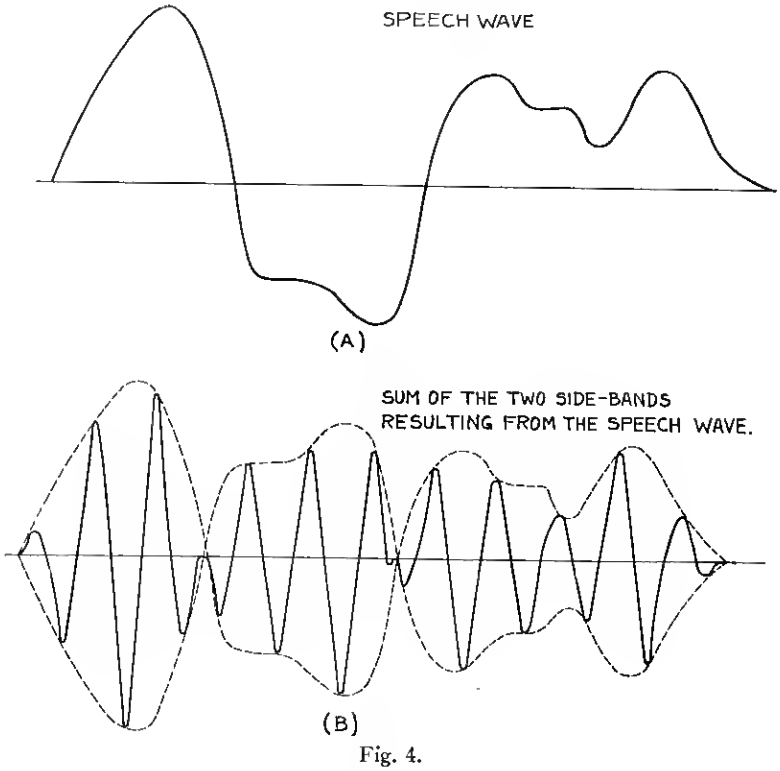

Fig. 4.

mission there will be no load on the amplifier and antenna except during actual speech, when it will be proportional to the intensity of the speech.

That these intermittent pulses of carrier frequency produced by a balanced modulator are equivalent to a modulated wave from which the carrier frequency component has been removed, may be easily shown. Consider a single sinusoidal component $S \cos (q\,t+\Theta)$, of the signaling wave which is applied to the balanced modulator. The resulting output current is a wave of carrier frequency, the amplitude of which is proportional to C and to S and varies cyclically with a frequency $\frac{q}{2\pi}$ between the values $+K\ C\ S$ and $-K\ C\ S$, where K is a

constant of proportionality and the negative amplitude indicates a reversal of phase during half of the audio frequency cycle. Such a variation may be represented by the expression

$$i = K\ C\ S \cos\ (q\ t + \theta) \cos\ p\ t. \qquad (9)$$

Taking the sum of these expressions for all the components of the signal gives the second term of equation (3) which was shown to represent the side-bands.

In estimating the power saved by carrier suppression the comparison should be made with a system transmitting the carrier which has been so adjusted that the power is used to the best advantage. So far as a single signal component is concerned this would call for making the carrier and side-band equal, as their product would then be a maximum. This, however, would imply that the distorting currents from the interaction of two side-band components would be as large as the signal currents themselves. That is to say, quality considerations require that the major part of the transmitted power be in the carrier component. Quantitative data on the relation between the ratio of carrier to side-band and the quality of transmission has been secured in the laboratories of the Western Electric Company, and it is hoped it will be published in the near future. Briefly, the results indicate that the good quality which is obtained when the carrier component is large falls off very rapidly as the magnitude of the carrier component is reduced so as to approach that of the side-band, the latter being measured when a sustained "ah" sound is used as the signal. Under these conditions the side-band is sustained at a value about equal to the maximum occurring in ordinary speech. That is to say, even the peak power in a carrier suppression system is less than the carrier component alone in an ordinary system adjusted to give the same side-band. From these considerations it appears that there has been a tendency to attach undue significance to "complete modulation," as a more or less unique and ideal condition of operation. For nothing revolutionary occurs as the carrier is decreased thru the value corresponding to that condition. The distortion due to interaction between the side-bands is present for larger values of carrier and continues to increase progressively for smaller values. The exact degree of modulation to be permitted therefore depends upon the standard of quality to be met. In a carrier suppression system the degree of modulation, k, approaches infinity more or less closely depending on the completeness of the suppression.

In addition to making possible the use of carrier suppression, homodyne reception presents other advantages. It furnishes a ready

means of increasing the intensity of the reproduced signal, since this is proportional to the carrier component at the receiver as well as to the side-band. Also, by making the carrier large, k is made very small and the distorting currents due to interaction of the side-bands become negligible. The use of a large local carrier in homodyne radio telephony assists in frequency selection in the same way as does the hetrodyne wave in radio telegraph reception. Suppose an interfering message is separated from the desired one by only a few thousand cycles and so is not entirely suppressed by the receiving selective circuits. Currents of voice frequency can be reproduced from its side-bands only by interaction with its own carrier, and hence they will be small compared with those of the desired message, which are proportional to the local carrier. On the other hand, the large currents due to the interfering message and local carrier will all have frequencies above the voice range, and so can be suppressed by selective circuits in the output of the detector.

The same general reasoning applies also to static interference. Appreciable interfering currents of signal frequency can result only from those components of the static wave which lie in the frequency range of the side-bands. Moreover, they will bear the same ratio to the signal currents as do the static components to the side-band components. We may conclude, then, that when means are provided for eliminating all of the static except that which is inherently inseparable from the signal, the disturbing effect of the residue is determined solely by the relative magnitude of the *side-band* components and the static components which lie in the same frequency range. As the object of high power stations is to make the signals large compared with the static, the importance of concentrating the power in the side-bands rather than in the carrier is obvious.

Effect of Radio Distortion

Let us pass now from the intentional modifications of a modulated wave and consider the effects of unintentional distortions. Limiting our attention first to systems in which the carrier is transmitted, we have to consider the effect of distortion such as might be introduced by the sending and receiving circuits and the transmitting medium. Assuming the characteristics of these to be known in terms of their transmission of sinusoidal components of various radio frequencies, we wish to determine their effect on the amplitudes and phases of the components of the reproduced wave. We shall assume the current-voltage relations in the transmission system to be linear, so

that no new frequencies are introduced. Then any possible distortion in the modulated wave may be represented by assigning the proper amplitudes and phases to all of the components. Corresponding to a single component of the signal we may write for the received wave

$$m = B\cos(pt-\phi)+B_+\cos[(p+q)\,t+\theta-\phi_+]+ B_-\cos[(p-q)\,t-\theta-\phi_-] \tag{10}$$

where the amplitude, B, and phase lag, ϕ, may vary in any arbitrary manner for the different components of the modulated wave. We shall assume that B is always large enough compared with B_+ and B_- that the interaction between the side-band components may be neglected. It will be seen that the single frequency components reproduced from the two side-bands are not in general equal nor in phase and may either aid or tend to neutralize each other. They will be of the form,

$$r = B\{B_+\cos[q\,t+\theta-(\phi_+-\phi)]+B_-\cos[q\,t+\theta-(\phi-\phi_-)]\}. \tag{11}$$

Taking the resultant of these two gives as the component of the reproduced wave,

$$r = R\cos(qt+\theta-\Psi) \tag{12}$$

where

$$R = B\sqrt{B_+^2+B_-^2+2\,B_+B_-\cos[(\phi_+-\phi)-(\phi-\phi_-)]} \tag{13}$$

$$\tan\Psi = \frac{B_+\sin(\phi_+-\phi)+B_-\sin(\phi-\phi_-)}{B_+\cos(\phi_+-\phi)+B_-\cos(\phi-\phi_-)}. \tag{14}$$

It is evident that both the amplitude, R, and the phase shift, Ψ, of the reproduced component depend upon both the amplitudes and phases of the corresponding components of both side-bands and on the phase of the carrier. The amplitude depends also on the amplitude B of the carrier, but as variations in this affect all components alike, they do not alter the wave form of the reproduced signal, but only its magnitude.

The expressions for the reproduced wave become much simpler for a system in which one side-band, say the lower, is suppressed. Then

$$B_- = 0 \tag{15}$$

and equations (13) and (14) reduce to

$$R = B\,B_+ \tag{16}$$

$$\Psi = \phi_+ - \phi. \tag{17}$$

The amplitude of the reproduced component is independent of the

phases in the modulated wave, and is proportional to the amplitude of the side-band component. Hence amplitude distortion of the reproduced wave can result only from unequal transmission of the different component frequencies of the side-band. The change in

AMPLITUDE-FREQUENCY AND PHASE-FREQUENCY CHARACTERISTICS FOR BAND-PASS FILTER.

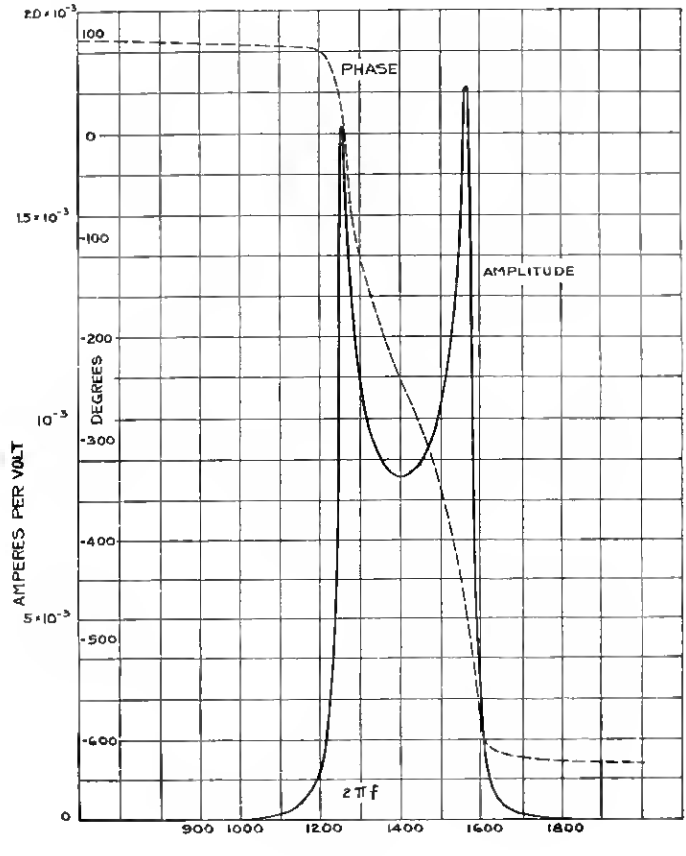

Fig. 5

the amplitude curve for the signal will be identical with that of the side-band. The phase shift, Ψ is independent of amplitude distortion of the modulated wave. It is equal to the difference between the phase lags of the side-band component and the carrier. Fortu-

nately the quality of telephone reproduction is not seriously impaired by shifting the phases of the various components by even as much as several cycles. In telegraphy, however, the shape of the signal current which operates the relay depends very much on the preservation of the proper phase relations of the components, and the entire nature of the signal may be changed by phase shifts of even a fraction of a cycle.

It is worth while then to examine some of the phase shifts which are likely to occur in practice. Transmission of a sinusoidal wave thru the free ether involves a phase lag proportional to the distance and to the frequency. Hence the phase lags, ϕ_+ and ϕ, due to this cause, will be proportional to $p+q$ and p respectively, and their difference, Ψ, will be proportional to q. Replacing Ψ by $h\,q$ in equation (12) and regrouping terms gives

$$r = R\cos\,[q(t-h)+\theta]. \tag{18}$$

By displacing the origin of time by h this becomes identical with the original signal component. Also, since h is independent of q, the same time shift brings all the components into agreement; that is, a phase shift proportional to the frequency does not distort the wave, but merely delays it by the corresponding time of transmission.[6] In considering the terminal circuits then it is only the departure of their phase lag versus frequency curve from a straight line that need be considered as a source of distortion. It is of interest to note here that for most filters this relation is approximately linear thruout the range of free transmission. The actual curves for a particular band filter are shown in Fig. 5, where there is plotted against frequency the relation of the amplitude and phase of the current at the third section of an infinite filter to those of the voltage applied to the first section. It will be noticed how the phase curve departs abruptly from a straight line at the edges of the band where the sudden drop in the amplitude curve occurs. Similarly Fig. 6 shows how in the current-voltage relation of a simple resonant circuit, the distortion of phase and of amplitude occur together.

In case both side-bands are transmitted, a simple relation is found if the distortion is symmetrical with respect to the carrier frequency. By this is meant that, however, the different components are distorted relative to each other, for every signal component the two corresponding side-band components are equal in amplitude and are

[6] For a fuller discussion of this point see a paper by T. C. Fry on "Theorie des binauralen Hörens nebst einer Erklärung der empirischen Hornbostel-Wertheimerschen Konstanten," "Physikalische Zeitschrift," 23, page 273, 1922.

shifted in phase relative to the carrier by the same amount; that is to say, for every value of q considered separately,

$$B_+ = B_- = B_\pm \tag{19}$$

$$\phi_+ - \phi = \phi - \phi_- = \delta. \tag{20}$$

Then

$$R = 2\,B\,B_\pm \tag{21}$$

$$\Psi = \delta. \tag{22}$$

The same considerations as to distortion of the side-band apply here as for the single side-band. There is one point, however, of some practical significance. As is evident from Fig. 6, the charac-

AMPLITUDE-FREQUENCY AND PHASE FREQUENCY CHARACTERISTICS OF SIMPLE RESONANT CIRCUIT.

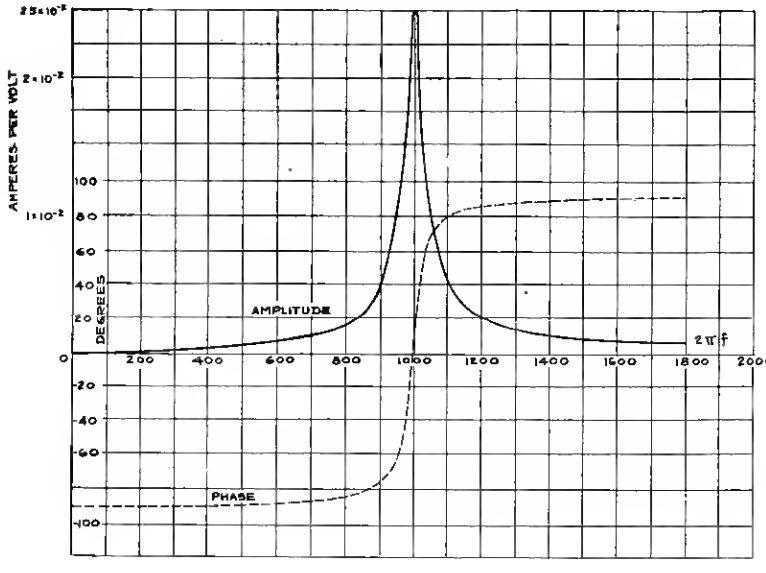

Fig. 6

teristics of a resonant circuit come very close to satisfying these symmetrical conditions if the carrier coincides with its resonance frequency. Its effect on the amplitude and phase curves of the reproduced signal may therefore be derived independently from the amplitude and phase curves of the tuned circuit. If, however, the

circuit be detuned, this symmetry is upset and we are forced to the complicated relations of equations (13) and (14), from which it is difficult to draw general conclusions.

An interesting case of unsymmetrical transmission is that in which one side-band is only partially suppressed owing to insufficient selectivity. Let us assume that the upper side-band is transmitted without distortion. Then for a given amplitude of the lower side-band its effect on the amplitude curve of the reproduced signal will be worst when the phase relations are such that for some frequencies it aids the upper side-band and for others it opposes. The greatest fractional change in the amplitude of any one signal component due to the presence of the lower side-band occurs when the two oppose. It is then reduced in the ratio

$$\frac{B_+ - B_-}{B_+} = 1 - \frac{B_-}{B_+}. \tag{23}$$

Thus, if the lower side-band component were a tenth of the upper, the most it could do would be to change the amplitude of the signal component by a tenth. Such a change would have little effect on telephone quality, particularly as it would have this maximum value at only a few frequencies. The case of telegraphy is rather different. Here the two side-bands lie so close together that it is practically impossible to separate them by radio selective circuits, and even when, as in wire transmission, so low a carrier frequency is used that the two may be separated fairly well by filters, the side-bands corresponding to the lower signal components differ so little in frequency that even a sharp filter does not produce very great discrimination between them. This, coupled with the fact that the phase shift of a filter ceases to be linear near the edge of the transmitted band, leads to very considerable amplitude distortion.

The effect of the unsuppressed side-band on the phases of the reproduced components is also rather complicated. Equation (14) shows that it is a maximum when

$$\begin{aligned} \phi_+ - \phi &= 0 \\ \phi - \phi_- &= 90°, \end{aligned} \tag{24}$$

in which case the presence of the lower side-band changes $tan\ \Psi$ from 0 to $\frac{B_-}{B_+}$. As in the case of amplitude, this effect is unimportant in telephony, but would need to be considered in telegraphy.

Phase of the Local Carrier

Coming now to homodyne reception, the important new factor to be considered is the fact that the carrier component is now perfectly arbitrary in amplitude and phase. This is true even tho the sending carrier is not suppressed, for, by suitably choosing the local carrier, the resultant of the two may be given any desired value. Since the amplitude of the carrier affects only the magnitude of the reproduced signal as a whole, we need consider here only the effect of arbitrary values of its phase, ϕ. For simplicity we shall assume that the modulated wave reaches the receiver unchanged except for the phase lags involved in undistorted transmission. Let us designate by ϕ_1 the phase lag of the carrier which is received from the transmitting station or would be received if it were not suppressed. Then

$$\phi = \phi_1 + \eta, \tag{25}$$

where η may be regarded as the phase displacement of the local carrier.

Consider first a system in which one side-band is suppressed. From equation (16), the amplitudes of the reproduced signal componants are independent of the phase of the carrier. From equation (17), the phase,

$$\Psi = \phi_+ - \phi_1 - \eta. \tag{26}$$

But $\phi_+ - \phi_1$ represents only the phase shifts of undistorted transmission; that is, the delay suffered by the signal as a whole. Hence the net result is that all components have their phases shifted by the same amount; namely, the phase displacement of the carrier, which can never be more than a single cycle. For telephony this is of no practical importance, but it is evident that in a telegraph system using side-band suppression and homodyne reception the phase of the local carrier would have to be very carefully controlled.

Consider now the case of homodyne reception of both side-bands received without distortion; that is,

$$B_+ = B_- = B_\pm \tag{27}$$

and

$$\phi_+ - \phi_1 = \phi_1 - \phi_- = h\,q. \tag{28}$$

From these relations and equations (13) and (14) we get

$$R = 2\,B\,B_\pm \cos\,\eta. \tag{29}$$

$$\Psi = h\,q. \tag{30}$$

This shows that the amplitude of every component varies as $\cos\,\eta$; that is, when the local carrier is in phase or 180° out of phase with the received carrier the reproduced components are maximum; for inter-

mediate values they decrease, becoming zero when the two carriers are in phase quadrature. The phase lag, $h\,q$, is that due to transmission alone; that is, the phases of the reproduced components are independent of the phase of the local carrier. Since the phase of the local carrier affects only the amplitudes and affects these the same for all components, it does not alter the wave form of the reproduced signal, but does affect its magnitude very materially.

Thus in a carrier telephone system fluctuations in the phase of the local carrier are much more serious when both side-bands are transmitted than when one is suppressed, the only effect then being an unimportant phase distortion. In a carrier telegraph system, however, the amplitude fluctuations which occur when both side-bands are transmitted may not be particularly troublesome since telegraph receiving apparatus is designed to operate over quite a range of signal intensity. The phase distortion occurring in single side-band transmission is however serious. It may perhaps be considered fortunate that the requirements as to phase regulation are least severe in telephony with a single side-band and in telegraphy with both side-bands, since these modes of operation appear on other grounds to be the most practical for the two cases.

In comparing single and double side-band transmission it is interesting to note that for equal sending power, the power of the reproduced signal component is twice as great with two side-bands as with one. However, the power of the same frequency resulting from static is also twice as great, so that the ratio of signal to interference is the same in both cases. To show this, let B_1 be the amplitude of a component of the single side-band and B_2 that of each of the corresponding components of the double side-band. Then equality of power gives

$$B_1^2 = 2\,B_2^2. \tag{31}$$

For the single side-band the amplitude of the reproduced component is

$$R_1 = g\,B_1, \tag{32}$$

where g is a constant of proportionality. The power,

$$P_1 = \tfrac{1}{2}\,g^2\,B_1^2 = g^2\,B_2^2. \tag{33}$$

(The resistance is here omitted as it is assumed constant thruout.) For the double side-band, since the two components are in phase, the resultant amplitude,

$$R_2 = 2\,g\,B_2, \tag{34}$$

and the power,

$$P_2 = 2\,g^2\,B_2^2 = 2\,P_1. \tag{35}$$

If the static be assumed to approximate an impulse, the amplitudes of all its components will be sensibly the same. If we call this amplitude S, then, in the case where the receiving circuit admits only one side-band, the amplitude of the reproduced interfering current of the frequency of the signal component is

$$I_1 = g\ S, \tag{36}$$

and its power,

$$W_1 = \tfrac{1}{2}\ g^2\ S^2. \tag{37}$$

With both side-bands this interfering current is made up of two equal components derived from the static components of frequencies $\frac{p+q}{2\pi}$ and $\frac{p-q}{2\pi}$ respectively. The phase difference ϵ between these two will be accidental, so for any one case the resultant amplitude,

$$I_2 = 2\ g\ S\ cos\ \epsilon, \tag{38}$$

and the power,

$$W_2 = 2\ g^2\ S^2\ cos^2\ \epsilon. \tag{39}$$

As all values of ϵ are equally probable, we may average W_2 with respect to ϵ, whence

$$\overline{W}_2 = g^2\ S^2 = 2\ W_1. \tag{40}$$

There is then no choice between one and two side-bands on the basis of the ratio of signal to interference. With a single side-band, the major advantage of economy in frequency range is secured at the expense of the minor disadvantage that to give the same response the amplification of the receiving set must be greater by a factor of two in power, or about three miles of standard cable.

Use of Non-synchronous Local Carrier

In practice, however, unless the receiving carrier frequency is controlled by the same source as the sending carrier, it is rather difficult to maintain even the frequencies alike, to say nothing of the phases. Let us suppose that the local carrier is out of synchronism by a small amount, n. Consider first the simplest case where the carrier is suppressed and one side-band only is transmitted. The local carrier beating with each component of this side-band gives a component of normal amplitude, but of a frequency differing from that of the original signal component by n. That is, all the frequencies of the speech are raised or lowered by the same amount, n. This must alter

the wave form very decidedly, but the surprising thing is that in telephony the intelligibility is not seriously affected when the difference is made as much as fifty cycles or so. The apparent pitch of the voice changes, of course, as n is varied.

If the carrier is transmitted, either intentionally or thru incomplete suppression, the situation is less favorable to asynchronous reception. The two carriers then beat together, giving a component of frequency n which may be troublesome if the received carrier is large. However, its frequency is generally below the voice range, and so it can be suppressed by a filter in the detector output. In addition the received carrier beating with the side-bands gives the components of the original signal. These are superposed on the displaced speech from the local carrier, the corresponding components of the two differing in frequency by n. As a result, the two sounds beat together just as two tuning forks would. For very little differences in frequency a periodic rise and fall in intensity is heard. When the difference is increased so that the individual beats can no longer be distinguished, a sensation of roughness results. And when the difference is made still greater the two waves may be heard as separate sounds of noticeably different pitch. The prominence of this beating effect depends, of course, upon the relative magnitude of the two carriers, since the two sets of speech currents are in the same ratio as the two carriers.

This effect of the received carrier may be very much reduced, and in the ideal case entirely eliminated, by the use of a balanced detector similar in structure to the balanced modulator of Fig. 3. It can be shown that with such a circuit the combination frequencies resulting from any two components applied at S are neutralized in the output circuit, while the combination of each with the carrier applied at C is transmitted. Thus if the side-band and received carrier enter together at S, the components having the original signal frequencies are eliminated and only the displaced components remain.

When the other side-band is added, the situation is still further complicated. In the absence of received carrier, the local carrier and one side-band give a set of components the frequencies of which are greater than those of the signal by n, while the carrier and other side-band give a set less by the same amount. These two sets combine in much the same way as do the displaced and normal speech obtained with a single side-band and received carrier. Here, however, the beat frequency is $2\ n$. Also, as the two sets are equal in amplitude, the beats will be much more pronounced, the intensity falling to zero each time the two waves are in opposition. For slow beats the

apparent pitch of the sound is half way between the frequencies of the two equal components, and so the normal voice frequency will be heard. With large frequency displacements of the carrier the two displaced speech waves, being of equal intensity, will be more easily distinguished than in the case of the single side-band.

It is interesting to note that this result, as well as the frequency shift that occurs with a single side-band, follows directly from the relations arrived at above for the phase displacement of a synchronous carrier. The non-synchronous carrier may be thought of as a synchronous one the phase of which is varied with the frequency of the departure from synchronism. With a single side-band it was shown that a phase displacement of the carrier affects only the phase of the reproduced component and that it changes this by an amount equal to its own displacement. This progressive phase displacement in all the components of the reproduced wave is, of course, equivalent to a change in their frequencies equal to the frequency displacement of the carrier. With both side-bands present, a phase displacement was shown to have no effect on the phases but to change the amplitudes of the reproduced components by the factor $\cos \eta$. Thus a progressive change in η will cause a cyclic variation in amplitude having two minima for each cycle of η; that is, a frequency of $2n$.

If the two side-bands are accompanied by the carrier there are added the beat note of the two carriers and the components of the original signal. The addition of a small amount of this speech of uniform amplitude to that of varying amplitude already present merely tends to make the variation slightly less pronounced. From the foregoing it appears that for telephony the most favorable condition for using a local carrier which is out of synchronism is that in which only one side-band is transmitted. Fairly considerable frequency variations are then permissible and asynchronous operation appears to have practical possibilities.

For telegraphy the case is quite different. In the first place the important components of telegraph signals are much lower in frequency, so that the side-band lies closer to the carrier, and a much smaller absolute displacement of the carrier frequency is needed to give the same effect as in the telephone case. Considering only such small displacements, it appears that the general addition or subtraction of frequencies which occurs with a single side-band will alter the shape of the signals quite seriously. The slow fluctuations in signal intensity which occur with both side-bands are probably less serious over most of the cycle. However, they might well cause some signals to be lost entirely each time the intensity passed thru zero.

For asynchronous reception, then, just as for the case of a local carrier displaced in phase, single side-band transmission is preferable for telephony and double side-band for telegraphy. The difficulties are of the same general nature in the two cases, but with the asynchronous carrier they are considerably greater. This is in agreement with the idea expressed above, that lack of synchronism may be looked upon as an aggravated case of phase displacement.